

**GENETIC VARIATIONS OF SIX ORANG ASLI
SUB-GROUPS IN PENINSULAR MALAYSIA
USING AUTOSOMAL AND X-CHROMOSOME
SHORT TANDEM REPEATS (STRs) ANALYSIS**

NUR AZEELAH BINTI ABDULLAH

UNIVERSITI SAINS MALAYSIA

2018

**GENETIC VARIATIONS OF SIX ORANG ASLI
SUB-GROUPS IN PENINSULAR MALAYSIA
USING AUTOSOMAL AND X-CHROMOSOME
SHORT TANDEM REPEATS (STRs) ANALYSIS**

by

NUR AZEELAH BINTI ABDULLAH

Thesis submitted in fulfillment of the requirements

for the degree of

Master of Science

March 2018

ACKNOWLEDGEMENT

In the name of ALLAH, the Most Beneficent, the Most Merciful.

This thesis would not have been possible without the support of many people. I am using this opportunity to express my gratitude to my main supervisor Associate Professor Dr. Zafarina Zainuddin for the encouragement, support, friendly advice and guidance during my research.

I would also like to thank the Director General of Department of Chemistry Malaysia for the permission to pursue my study. Not forgetting the Dean of School of Health Sciences and all the laboratory staff, who provided me with the facilities required and conducive condition for my research. Deepest gratitude to the members of the Forensic Science postgraduate for sharing the literature and invaluable assistance.

Finally, I also wish to express my love and gratitude to my beloved husband, my little princess; Anis, Alia, Amni and families for their understanding and support through the journey of my studies.

TABLE OF CONTENT

Acknowledgement	ii
Table of Content	iii
List of Tables.	viii
List of Figures	xiii
List of Symbols, Abbreviations and Acronyms	xv
Abstrak.....	xvii
Abstract.....	xix
 CHAPTER 1 : INTRODUCTION	 1
1.1 Human Genome	1
1.1.1 Nuclear DNA	1
1.1.2 Mitochondrial DNA	2
1.2 Deoxyribonucleic Acid (DNA) Profiling.....	4
1.3 DNA Typing Methods	4
1.3.1 Restriction Fragment Length Polymorphism (RFLP).....	5
1.3.2 Polymerase Chain Reaction (PCR).....	5
1.3.2(a) HLA-DQ α and PolyMarkers	6
1.3.2(b) Short Tandem Repeats (STR).....	7
1.4 Population Studies	12
1.4.1 The Orang Asli Population in Peninsular Malaysia.....	12

1.4.1(a)	Semang	16
1.4.1(b)	Senoi	16
1.4.1(c)	Proto-Malay	17
1.4.2	Population Genetics	18
1.4.3	Population Database.....	19
1.5	Objectives of Study	20
CHAPTER 2 : MATERIAL AND METHOD		21
2.1	Biological Samples	21
2.2	Materials and Reagents	24
2.2.1	DNA Extraction Kit	24
2.2.2	Human Polymerase Chain Reaction (PCR) Amplification Kit.....	24
2.2.2(a)	AmpF/STR® Identifiler® Direct PCR Amplification Kit	24
2.2.2(b)	Investigator® Argus X-12 QS Kit.....	26
2.2.3	Matrix Standard Kit	29
2.2.4	DNA Size Standard Kit.....	29
2.2.5	Reagents on the Applied Biosystems® 3130xl Genetic Analyzer.....	29
2.3	Instrumentation and Apparatus	30
2.4	Methods	32
2.4.1	Contamination Prevention	32
2.4.2	DNA Extraction	32
2.4.3	Determination of Genomic DNA Concentration	33

2.4.4	Polymerase Chain Reaction (PCR) Amplification	33
2.4.4(a)	AmpF/STR [®] Identifier [®] Direct PCR for the Autosomal STRs	34
2.4.4(b)	Investigator [®] Argus X-12 QS for X-chromosome STR.....	34
2.4.5	Electrophoresis on the Applied Biosystems [®] 3130xl Genetic Analyzer	37
2.4.5(a)	Spectral Calibration.....	37
2.4.5(b)	Analysis of the Amplified Products	37
2.4.6	GeneMapper [®] ID Software Analysis	40
2.5	Statistical Analysis Method	41
2.5.1	Common Statistical Analysis Method for Autosomal STR and X- STR	41
2.5.1(a)	Allele Frequency	41
2.5.1(b)	Gene Diversity (GD)	42
2.5.1(c)	Hardy-Weinberg Equilibrium (HWE).....	43
2.5.1(d)	Polymorphism Information Content (PIC).....	43
2.5.1(e)	Power of Discrimination (PD).....	44
2.5.1(f)	Power of Exclusion (PE)	45
2.5.1(g)	Analysis of Molecular Variance (AMOVA)	45
2.5.1(h)	F-statistics; Inbreeding Coefficient (F_{IS}), Population Fixation Index (F_{ST}) and Overall Fixation Index (F_{IT}).....	47
2.5.1(i)	Principal Component Analysis (PCA)	49

2.5.1(j) Phylogenetic Tree	49
2.5.2 Specific Statistical Analysis Method for X-STR	50
2.5.2(a) Mean Exclusion Chance (MEC).....	50
CHAPTER 3 : RESULTS	51
3.1 Genomic DNA	51
3.2 Allele Frequency for 15 Autosomal STR Loci	51
3.3 Allele Frequency for 12 X-STR Loci	65
3.4 Gene Diversity (GD).....	80
3.5 Hardy-Weinberg Equilibrium (HWE)	82
3.6 Polymorphism Information Content (PIC)	85
3.7 Power of Discrimination (PD).....	87
3.8 Power of Exclusion (PE).....	90
3.9 Analysis of Molecular Variance (AMOVA).....	92
3.10 F-statistics	94
3.11 Principal Component Analysis (PCA)	95
3.12 Phylogenetic Tree	98
3.13 Mean Exclusion Chance (MEC)	101
CHAPTER 4 : DISCUSSION.....	104
4.1 Genetic Variation in the Orang Asli Populations.....	104
4.2 Forensic Efficiency Parameter.....	105

4.3	Relationship between the Orang Asli Sub-groups and other world populations.....	107
	CONCLUSION.....	109
	REFERENCES	111
APPENDICES		
Appendix A	: Informed consent	
Appendix B	: Questionnaire	
Appendix C	: Human ethical approval	
Appendix D	: Concentration and purity of genomic DNA	
Appendix E	: Results of 15 autosomal STR typing of six Orang Asli sub-groups in Peninsular Malaysia	
Appendix F	: Results of 12 X-STR typing of six Orang Asli sub-groups in Peninsular Malaysia	

LIST OF TABLES

	Page
Table 1.1 : Groups and sub-groups of Orang Asli in Peninsular Malaysia	15
Table 2.1 : The sampling location and number of samples for six Orang Asli sub-groups	22
Table 2.2 : The AmpFISTR® Identifiler® Direct PCR Amplification kit contents	25
Table 2.3 : The amplified loci, locations, dyes, the allele contained in the allelic ladder and the genotype of the control DNA in the Identifiler Direct kit.....	25
Table 2.4 : The Investigator® Argus X-12 QS Kit contents.....	27
Table 2.5 : Linkage group for 12 X-chromosome STR.....	27
Table 2.6 : The amplified loci, locations, dyes, the allele contained in the allelic ladder and the genotype of the control DNA in the Investigator Argus X-12 QS kit.....	28
Table 2.7 : List of the matrix standard kit.....	31
Table 2.8 : List of the DNA size standard kit.....	31
Table 2.9 : The amplification parameter for AmpFISTR Identifiler Direct PCR Amplification Kit.....	36

Table 2.10	: The amplification parameter for Investigator Argus X-12 QS Amplification Kit.....	36
Table 2.11	: Run module setting for AmpFlSTR® Identifiler® Direct PCR and Investigator® Argus X-12 QS.....	38
Table 2.12	: Sample preparation.....	40
Table 3.1	: Allele frequencies of 15 autosomal STRs for the Semai sub-group (N=42).....	52
Table 3.2	: Allele frequencies of 15 autosomal STRs for the Che Wong sub-group (N=28).....	54
Table 3.3	: Allele frequencies of 15 autosomal STRs for the Orang Kanaq sub-group (N=11).....	56
Table 3.4	: Allele frequencies of 15 autosomal STRs for the Lanoh sub-group (N=26).....	57
Table 3.5	: Allele frequencies of 15 autosomal STRs for the Bateq sub-group (N=25).....	59
Table 3.6	: Allele frequencies of 15 autosomal STRs for the Kensi sub-group (N=32).....	61
Table 3.7	: Allele frequencies of 15 autosomal STRs for the combined Orang Asli populations (COAP) (N=164).....	63

Table 3.8	: Allele frequencies of 12 X-STRs for the Semai sub-group (N=42).....	66
Table 3.9	: Allele frequencies of 12 X-STRs for the Che Wong sub-group (N=28).....	68
Table 3.10	: Allele frequencies of 12 X-STRs for the Orang Kanaq sub-group (N=11).....	70
Table 3.11	: Allele frequencies of 12 X-STRs for the Lanoh sub-group (N=26).....	71
Table 3.12	: Allele frequencies of 12 X-STRs for the Bateq sub-group (N=25).....	73
Table 3.13	: Allele frequencies of 12 X-STR for the Kensiu sub-group (N=32).....	75
Table 3.14	: Allele frequencies of 12 X-STRs for the combined Orang Asli populations (COAP) (N=164).....	77
Table 3.15	: Gene diversity values of 15 autosomal STR genetic loci for the six Orang Asli sub-groups and combined Orang Asli populations (COAP).....	81
Table 3.16	: Gene diversity values of 12 X-STR genetic loci for the six Orang Asli sub-groups and combined Orang Asli populations (COAP).....	81

Table 3.17	: The exact test of HWE of the 15 autosomal STR loci calculated for Semai, Che Wong and Orang Kanaq sub-groups.....	83
Table 3.18	: The exact test of HWE of the 15 autosomal STR loci calculated for Lanoh, Bateq and Kensiu sub-groups.....	83
Table 3.19	: The exact test of HWE of the 12 X-STR loci calculated for Semai, Che Wong and Orang Kanaq sub-groups.....	84
Table 3.20	: The exact test of HWE of the 12 X-STR loci calculated for Lanoh, Bateq and Kensiu sub-groups.....	84
Table 3.21	: The Polymorphism Information Content (PIC) of the 15 autosomal STR loci in the six Orang Asli sub-groups.....	86
Table 3.22	: The Polymorphism Information Content (PIC) of the 12 X-STR in the six Orang Asli sub-groups.....	86
Table 3.23	: The Power of Discrimination (PD) of the 15 autosomal STR in the six Orang Asli sub-groups.....	88
Table 3.24	: The Power of Discrimination for male (PD_M) of the 12 X-STR in the six Orang Asli sub-groups.....	89
Table 3.25	: The Power of Discrimination for female (PD_F) of the 12 X-STR in the six Orang Asli sub-groups.....	89

Table 3.26	: The Power of Exclusion (PE) of 15 autosomal STR in the six Orang Asli sub-groups.....	91
Table 3.27	: The Power of Exclusion (PE) of 12 X-STR in the six Orang Asli sub-groups.....	91
Table 3.28	: The AMOVA analysis for the 15 autosomal STR in the six Orang Asli sub-groups.....	93
Table 3.29	: The AMOVA analysis for the 12 X-STR in the six Orang Asli sub-groups.....	93
Table 3.30	: The MEC values of the 12 X-STR loci calculated for Semai and Che Wong sub-groups.....	102
Table 3.31	: The MEC values of the 12 X-STR loci calculated for Orang Kanaq and Bateq sub-groups.....	102
Table 3.32	: The MEC values of the 12 X-STR loci calculated for Bateq and Kensiu sub-groups.....	103

LIST OF FIGURES

	Page
Figure 1.1 : Human genome structure	3
Figure 1.2 : Human mitochondrial DNA.....	3
Figure 1.3 : The chromosomal location for DNA (STR) profiling.....	10
Figure 1.4 : Inheritance pattern for the X-chromosome	10
Figure 1.5 : The physical location of X-STR markers and the four proposed linkage groups	11
Figure 1.6 : Map showing the Orang Asli distribution in Peninsular Malaysia	14
Figure 2.1 : The schematic diagram of the methodology for the autosomal STR and X-STR analysis	23
Figure 3.1 : Principal Component Analysis (PCA) using allele frequencies of the 15 autosomal STR loci for the six Orang Asli sub-groups, COAP and other world populations	96
Figure 3.2 : Principal Component Analysis (PCA) using allele frequencies of the 12 X-STR loci for the six Orang Asli sub- groups, COAP and other world populations	97

Figure 3.3	:	Neighbor-Joining phylogram constructed based on genetic distances of the allele frequencies in autosomal STR for six Orang Asli sub-groups.....	99
Figure 3.4	:	Neighbor-Joining phylogram constructed based on genetic distances of the allele frequencies in autosomal STR for combined Orang Asli populations (COAP) with other world populations	100

LIST OF SYMBOLS, ABBREVIATIONS AND ACRONYMS

Θ	Theta
∞	Infinity
$^{\circ}\text{C}$	Degree Celcius
μl	microliter
AMOVA	Analysis of molecular variance
ATP	Adenosine Triphosphate
bp	base pair
CE	Capillary Electrophoresis
COAP	Combined Orang Asli population
DNA	Deoxyribonucleic Acid
EDTA	Ethylenediaminetetraacetic acid
F_{IS}	Inbreeding coefficient
F_{IT}	Overall fixation index
F_{ST}	Population fixation index
g	gram
GD	Gene Diversity
HET	Heterozygosity
HLA	Human Leukocyte Antigen
HV	Hypervariable region
HWE	Hardy-Weinberg Equilibrium
KV	Kilovolt
LE	Linkage Equilibrium
min	minute

ml	milliliter
mtDNA	Mitochondrial DNA
ng	nanogram
NRC	National Research Council
PCA	Principal Component analysis
PCR	Polymerase Chain Reaction
PD	Power of Discrimination
PE	Power of Exclusion
PIC	Polymorphism Information Content
rCRS	Revised Cambridge Reference Sequence
RFLP	Restriction Fragment Length Polymorphism
RNA	Ribonucleic Acid
rpm	revolution per minute
sec	second
STR	Short Tandem Repeats
SWGDM	Scientific Working Group on DNA Analysis Methods
VNTR	Variable Number Tandem Repeats

**VARIASI GENETIK BAGI ENAM SUB-KUMPULAN ORANG ASLI DI
SEMENANJUNG MALAYSIA MENGGUNAKAN ANALISIS *SHORT
TANDEM REPEATS* (STRs) AUTOSOM DAN KROMOSOM-X**

ABSTRAK

Penduduk Orang Asli hanya mewakili kira-kira 0.5% daripada jumlah penduduk Malaysia. Terdapat 18 suku kaum Orang Asli yang diklasifikasikan di bawah Semang, Senoi dan Proto-Melayu. Mereka mempunyai pelbagai budaya, tradisi, kepercayaan dan bahasa, bergantung kepada etnik dan lokasi. Kajian ini melibatkan sebanyak 164 sampel darah dari enam suku kaum Orang Asli di Semenanjung Malaysia; Semai, Che Wong, Orang Kanaq, Lanoh, Bateq dan Kensiu menggunakan kaedah PCR bagi 15 lokus STR autosom menggunakan sistem *Identifiler Direct* dan 12 lokus X-kromosom STR menggunakan sistem *Argus X-12 QS*. Kekerapan alel dan beberapa parameter forensik untuk STR serta struktur genetik dalam kumpulan Orang Asli turut dianalisis. Semua lokus STR autosom bagi suku kaum Semai dan Lanoh adalah tidak menyimpang daripada keseimbangan Hardy-Weinberg (HWE), begitu juga dengan semua lokus bagi X-kromosom STR untuk suku kaum Che Wong. Gabungan kuasa nilai pengecualian (CPE) adalah lebih tinggi daripada 0.999. Gabungan kuasa nilai diskriminasi (CPD) juga menyokong teori bahawa tidak ada dua individu yang akan mempunyai genotip yang sama kecuali untuk kembar seiras. Penanda FGA adalah penanda yang paling bermaklumat untuk ujian forensik daripada 15 lokus STR autosom manakala penanda DXS10146 adalah penanda yang paling bermaklumat untuk sistem X-STR Orang Asli di Semenanjung Malaysia. Plot *Principal Coordinate Analysis* (PCA) yang dihasilkan dalam kajian ini

menunjukkan bahawa suku kaum Orang Asli dikelompokkan bersama yang menunjukkan kemungkinan mereka berasal dari keturunan yang sama. Ini menunjukkan keunikan suku kaum Orang Asli dan bukti pertalian genetik purba mereka kerana mereka mewakili gelombang penghijrahan manusia yang pertama keluar dari Afrika. Data profil STR dan X-STR ini adalah penting untuk aplikasi forensik dan penubuhan data penduduk Orang Asli di Semenanjung Malaysia di mana tiada pangkalan data STR tersedia untuk penduduk ini.

GENETIC VARIATIONS OF SIX ORANG ASLI SUB-GROUPS IN PENINSULAR MALAYSIA USING AUTOSOMAL AND X-CHROMOSOME SHORT TANDEM REPEATS (STRs) ANALYSIS

ABSTRACT

The Orang Asli population represent only about 0.5% of total population of Malaysia. There are 18 sub-groups of Orang Asli that belong to the Semang, Senoi and Proto-Malay. These sub-groups have diverse cultures, traditions, beliefs and languages depending on their ethnicity and location. In this study, a total of 164 blood samples from six Orang Asli sub-groups in Peninsular Malaysia; Semai, Che Wong, Orang Kanaq, Lanoh, Bateq and Kensiu were PCR-type for 15 autosomal STR loci using Identifiler Direct system and 12 X-chromosomal STR loci using Argus X-12 QS system. The allele frequencies and several forensic parameters for STR as well as the genetic structure in the Orang Asli sub-groups were determined. The agreement with Hardy-weinberg Equilibrium (HWE) was confirmed for all loci in Semai and Lanoh sub-groups for autosomal STR results and Che Wong for X-STR results. The combined power of exclusion (CPE) were higher than 0.999. The combined power of discrimination (CPD) also supports the theory that no two individuals would have the same genotype except for identical twins. FGA is the most variable and informative marker for forensic testing of the 15 autosomal STR loci while DXS10146 is the most informative marker for X-STR system of the Orang Asli in Peninsular Malaysia. The Orang Asli sub-groups were plotted closed to each other in the Principal Component Analysis (PCA), suggesting some degree of common ancestry between the subgroups. These findings indicate the uniqueness of the Orang Asli sub-groups and the evidence of their ancient genetic affinities as they represent the first human

migration wave out of Africa. The STR and X-STR profile dataset are important for forensic applications and for establishment of the Orang Asli population data in Peninsular Malaysia, which are currently not available for this population.

CHAPTER 1

INTRODUCTION

1.1 Human Genome

The human genome (Figure 1.1) consist over 3 billion base pairs. The genome is a complete set of nucleic acid sequence for humans that organized into 23 paired chromosomes located in the cell nucleus. The genome also includes the mitochondrial DNA, a comparatively small circular molecule present in each mitochondrion. Human genomes include both protein-coding DNA genes and noncoding DNA (Butler, 2010).

1.1.1 Nuclear DNA

The nuclear DNA is arranged on 23 pairs of chromosomes. Twenty two (22) pairs are autosome and the other pair is sex-determining chromosome, with XX for female and XY for male individuals. DNA materials comprise coding regions (genes) that code for protein and non-coding regions that do not code for information required for protein synthesis. The non-coding regions make up about 90% of the genome and referred as 'junk' DNA. A part of this non-coding DNA comprises repetitive sequences and highly polymorphic. The location of a gene in the non-coding region known as locus or loci, which have been used for characterization and mapping of a particular region of human chromosome (Panneerchelvam & Norazmi, 2003).

1.1.2 Mitochondrial DNA

Mitochondrial DNA (mtDNA) encode for cellular energy production is located in the mitochondria organel (Anderson *et al.*, 1981). It converts chemical energy from food into adenosine triphosphate (ATP) for metabolism. In humans, mtDNA is a double-stranded circular molecule of approximately 16,569 base pairs (Figure 1.2). These two strands of mtDNA are different in their nucleotide content, in which the light strand (L-strand) is cytosine-rich and heavy strand (H-strand) is guanine-rich.

mtDNA is maternally inherited and passed unchanged along from the mother to offspring (male/female) unless spontaneous mutation occur. The importance of mtDNA in human identification mainly involved the D-loop region which consists of three hypervariable regions known as hypervariable region1 (HV1); hypervariable region 2 (HV2) and hypervariable region 3 (HV3). These regions exhibit multiple variations between individuals (Anderson *et al.*, 1981; Haslindawaty *et al.*, 2010).

The first human mtDNA was sequenced by Sanger's laboratory in 1981 (Anderson *et al.*, 1981) and was later revised as Cambridge Reference Sequence (rCRS). Variations in human mtDNA are reported by comparing an individual's sequence to CRS. Human mtDNA can also be used to exclude possible matches between missing person and unidentified remains (Monsalve & Hagelberg, 1997). mtDNA is a better choice than nuclear DNA in certain cases because the greater number of copies of mtDNA per cell increase the chance of obtaining a useful sample (Parsons & Coble, 2001).

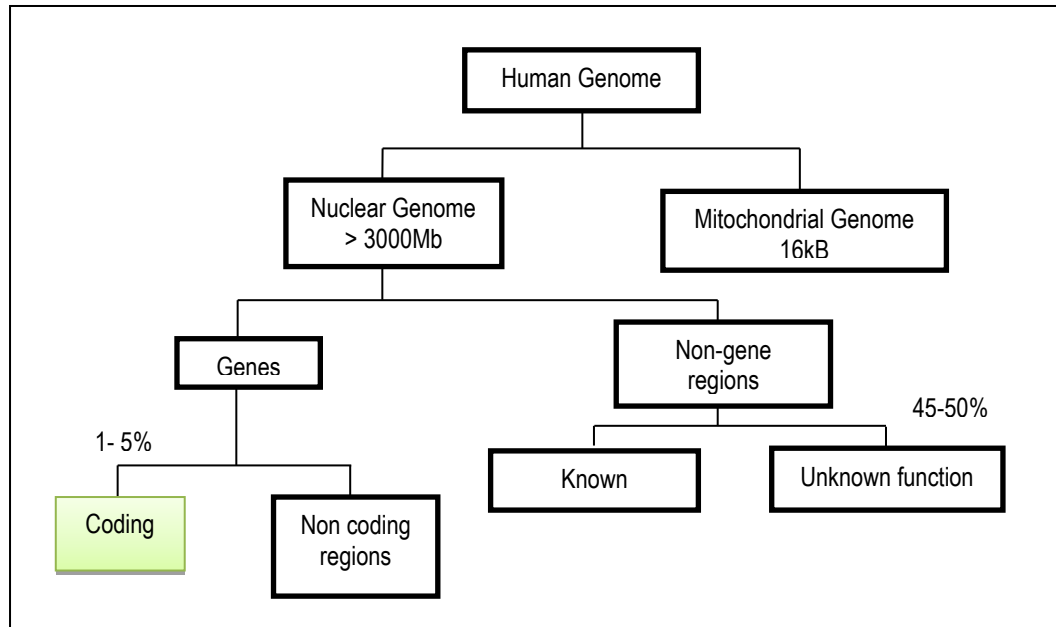


Figure 1.1: Human Genome Structure

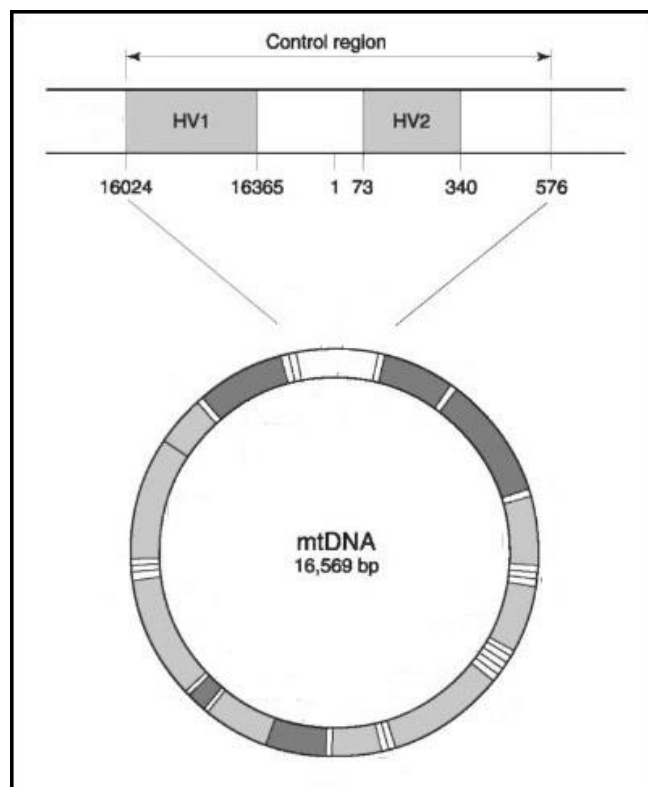


Figure 1.2: Human mitochondrial DNA

(Modified and adapted from Forensic DNA Typing, (2005), p124)

1.2 Deoxyribonucleic Acid (DNA) Profiling

DNA profiling is a technique to identify an individual's DNA characteristics, which is unique to each individual. This technique uses repeated DNA sequences that are highly variable and differ from each individual (Butler, 2006).

DNA profiling is commonly used in forensics to identify and link an individual to the evidence from a crime scene. It is also being used in parentage testing. The process of DNA profiling begins with the laboratory analysis of biological sample to determine their genetic marker type; then comparison of the genetic marker to determine the match and whether the sample could have come from the same source. Finally the statistical analysis of the population frequency to determine the probability that such a match might be observed by chance in comparison of samples from different persons (National Research Council (US) Committee on DNA Technology in Forensic Science, 1992). The development of technologies in DNA profiling and the ability to obtain genetic profile from minute amount of DNA samples become very useful in forensic science.

1.3 DNA Typing Methods

DNA typing was discovered by Professor Alec Jefferys and his colleagues in 1985. They found that certain regions of DNA contain sequences that are repeated over and over again and differ from each individual while studying the gene coding for myoglobin (Jefferys *et al.*, 1985). They used a specific probe to isolate a piece of

DNA and discovered that every analyzed sample resulted in numerous bands. The technique turned out to be very unique and enables human identification.

1.3.1 Restriction Fragment Length Polymorphism (RFLP)

RFLP technique is the first method used in DNA profiling and generates DNA fragments of different lengths (Botstein *et al.*, 1980). The DNA collected from cell is cut into small pieces using a restriction enzyme and generates thousands of DNA fragments with different sizes as consequence of variations between DNA sequences of different individuals. The resulting restriction fragments are then separated accordingly to their length by gel electrophoresis. The separated fragments are transferred to a membrane via Southern blot method and labeled using radioactive probe that are complementary to the sequence in the fragments (Kobilinsky *et al.*, 2005). The allele is called based on the fragment length. However, this technique is laborious and require large amount of good quality DNA. This early technique has now being replaced by PCR-based assay.

1.3.2 Polymerase Chain Reaction (PCR)

In 1983, Kary Mullis and members of the Human Genetics group have developed PCR technique to amplify a single copy of DNA to generate billions copies of a targeted sequence (Butler, 2006). This technique is easy, cheap and reliable for DNA amplification. It is suitable for forensic DNA analysis as it requires less amount of genomic DNA. Application of PCR technique in DNA profiling improves the

discriminating power and increases the possibility to recover DNA from limited amount or degraded samples.

1.3.2(a) HLA-DQ α and PolyMarkers

The first PCR typing system used in forensic analysis was HLA-DQ α . This technique is based on sequence variation at the DQ α locus of the human leukocyte antigen (HLA) gene found in chromosome 6. In this system, the amplified DNA reacts with a variety of probes that bind to a blot. Sequence specific probe for each allele will be immobilized on the blot strip. A chemical reaction causes the dot from colorless substrate to a blue precipitate. The pattern of dots corresponds to the alleles present and allows the DNA type of the sample to be determined.

The first DQ α kit detects six (6) common alleles in combination, and defined only 21 possible genotypes. Limited number of alleles present in the population and limited number of genotypes has resulted in much lower power of discrimination as compared to RFLP analysis. However, the HLA-DQ α was the best and rapid approach when dealing with very small amount of DNA samples (Rudin & Inman, 2002).

The AmpliType pm +DQA1 kit, also known as PolyMarker consist of five (5) genetic loci was introduced by Perkin-Elmer/Roche to meet the needs for a greater power of discrimination (Kobilinsky *et al.*, 2005). However, the results from this system were difficult to interpret when the samples contained more than one contributor and the power of discrimination is still low.

1.3.2(b) Short Tandem Repeats (STR)

STR has been widely used by forensic DNA community all over the world. This method is highly polymorphic with short repeated sequences of DNA in the range of 2 to 7 base pairs. STR markers are found in the genome in every 10,000 nucleotides (Edwards *et al.*, 1991). Their efficiency and hyper variability to amplification make them an ideal marker for the purpose of human identification.

The DNA marker used today for forensic DNA profiling is based on PCR amplification of STR. These STR loci are amplified using sequence-specific primers and the amplified products are analyzed using capillary electrophoresis (Butler, 2006). The STR repeat regions are classified into several groups according to the size of the repeat unit. They are known as dinucleotide, trinucleotides, tetranucleotide, pentanucleotide and hexanucleotide. The forensic DNA community has accepted and validated tetranucleotide repeats for use in the human identification (Butler, 2005).

Each STR allele is shared by 5 – 20% of individual and the genotype may identify an individual accurately. The power of STR analysis increased when multiple STR loci were examined together. Most of the forensic DNA kit is designed for multiplex systems which allow multiple STR loci to be amplified simultaneously. To avoid overlapping within the locus, different size range was selected. In STR analysis, an incomplete repeat unit is called microvariant allele and the availability of this microvariant will increase the power of discrimination significantly (Kobilinsky *et al.*, 2005). The selected STR loci should have high level of polymorphism and heterozygosity in order to discriminate between individuals.

Autosomal STR

Each individual inherited 22 pairs of autosomal chromosomes from biological mother and biological father, respectively. Therefore each person has a unique DNA except for identical twins (Butler, 2005). This unique characteristic of autosomal STR makes it as a powerful tool for many applications such as population genetic studies, relationship testing, identification testing and forensic matching identification.

The forensic DNA typing community leading by the Federal Bureau of Investigation (FBI) of the United State has developed technology for use in the human DNA identification. They have selected 13 STR loci (Figure 1.3) to generate a nationwide DNA database called the Combined DNA Index System (CODIS). The 13 CODIS loci are D8S1179, D21S211, D7S820, CSF1PO, D3S1358, TH01, D13S317, D16S539, vWA, TPOX, D18S51, D5S818 and FGA while 2 additional markers D2S1338 and D19S433 were added by the European DNA community. The available commercial STR kits have used these combinations of loci to become the core loci in forensic identification either in parentage testing or casework (Butler & Hill, 2012).

X-chromosome STR

X-STR is useful in certain forensic casework such as missing persons, criminal incest, immigration, deficiency paternity or other questioned relationships have special characteristics that make them particularly. The X chromosome contains 153 million base pairs with approximately 5 percent of total female DNA in cell. A male inherited X-chromosome only from the mother while a female inherited two X-chromosome

from father and mother, respectively (Figure 1.4). The X-STR markers can be used as a DNA complementary test to autosomal and Y-STRs especially when solving kinship deficiency cases (Szibor *et al*, 2003).

The X-STR markers (Figure 1.5) were divided into four linkage groups which yielded independent genotype information (Szibor, 2007). The commercial typing kit available in the market unlinked X-STR which are DXS8378, DXS7132, HPRTB and DXS7423 as the core X-STR loci. The latest commercial kit available is the Investigator Argus X-12 PCR Amplification QS Kit by Qiagen that amplified 12 X-STR loci namely DXS10103, DXS8378, DXS10101, DXS10134, DXS10074, DXS7132, DXS10135, DXS7423, DXS10146, DXS10079, HPRTB and DXS10148 (Shrivastava P. *et al*,. 2015).

The population genetics parameter useful in forensic science has to be calculated for X-STR evidence sample to illustrate the power of the analysis. Reinhard Szibor and his colleague from Institute of Rechtsmedizin, University of Magdeburg Germany have created a website to provide a database comprising published population X-STR data for populations from several countries. This website is known as Forensic ChrX Research website (ChrX-STR) and available online via <http://chrX-str.org>.

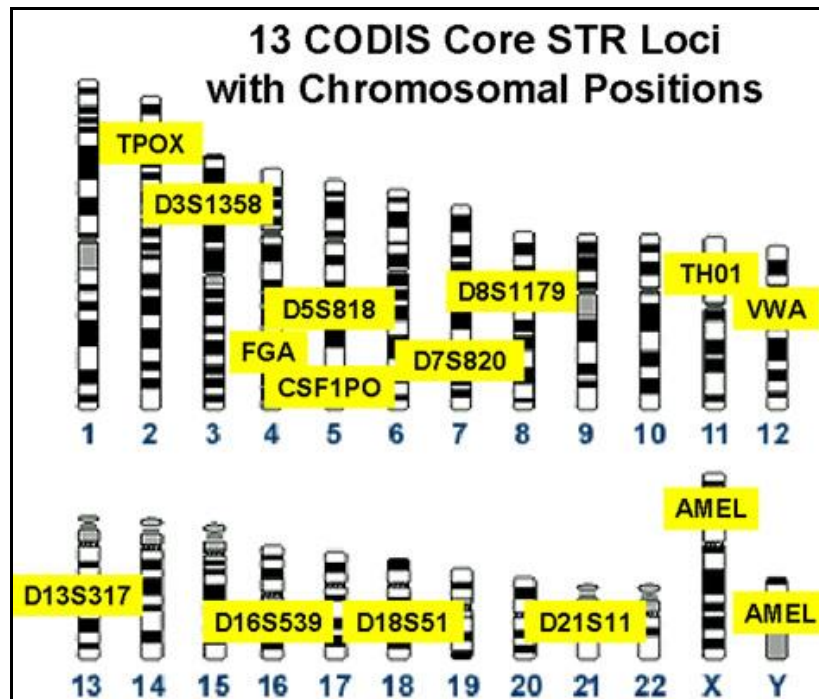


Figure 1.3: The chromosomal location for DNA (STR) profiling

(Source: <http://www.cstl.nist.gov/div831/strbase/fbicore>)

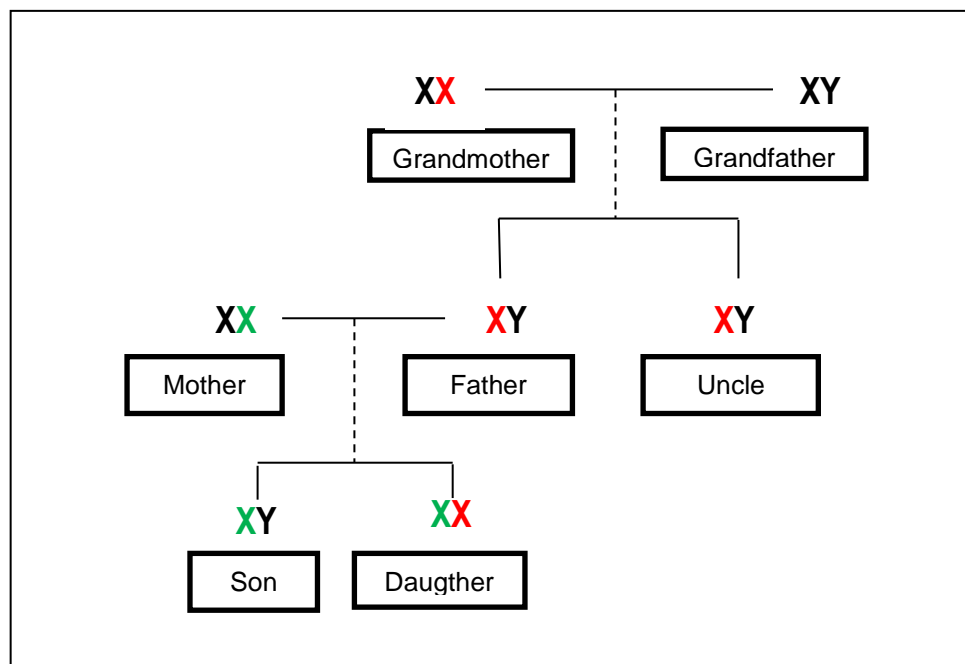


Figure 1.4: Inheritance pattern for the X-chromosome

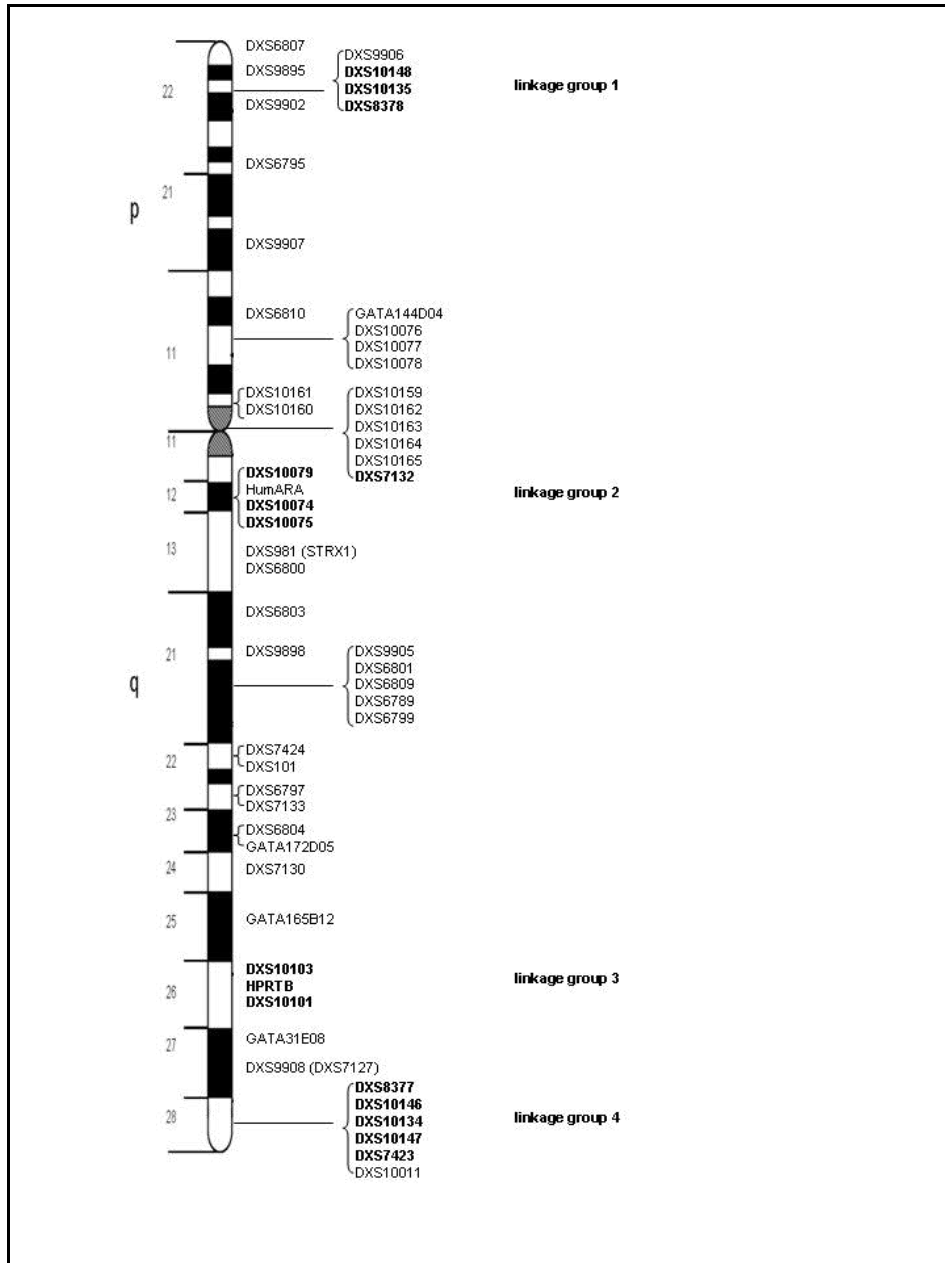


Figure 1.5: The physical location of X-STR markers and the four proposed linkage groups (Source: Investigator® Argus X-12 Handbook (June 2013), p7)

1.4 Population Studies

1.4.1 The Orang Asli Population in Peninsular Malaysia

Peninsular Malaysia or West Malaysia, consists of eleven states and two federal territories. It is separated from East Malaysia by the South China Sea. Population in Peninsular Malaysia comprises of various diversity of ethnicity that includes Malay, Indian, Chinese and the Orang Asli (Hood, 2006; Lye, 2001).

The indigenous people or Orang Asli is the earliest population arrived in Peninsular Malaysia about 50,000 years ago (Hill *et al.*, 2006). The distribution of Orang Asli in Peninsular Malaysia is shown in Figure 1.6. The Orang Asli is not a homogenous groups, each of them has its own culture and distinct language. The ‘Orang Asli’ in a Malay term also known as ‘original people’ or ‘first people’ with three main ethnic groups; Semang, Senoi and Proto Malay and each group consists of 6 sub-groups as listed in Table 1.1.

The route of Orang Asli’s migration into Peninsular Malaysia is quite complicated. Metspalu *et al.* (2006) reported, modern humans were first migrated across the Sahara, out of Egypt to the Levant through the northern route. The modern humans were also believed had left Africa via southern coastal route as a single group crossing the mouth of Red Sea from Eritrea to India, Southeast Asian (SEA) and subsequently reached the isolated Sahul continent (Oppenheimer, 2009). A resource-rich living environment offered in the coastal area making southern coastal route very likely path taken during the migration.

The human migration of Orang Asli via southern route is reasonable, as recent mtDNA study on relict populations of Southeast Asia, the Andaman and Nicobar Island have also concluded that major human dispersals were via southern route (Macaulay *et al.*, 2005; Thangaraj *et al.*, 2005). Numerous theories either based on linguistic, archeological evidence and historical as well as genetic studies have been proposed on the migrations route of Orang Asli in Peninsular Malaysia (Hill *et al.*, 2006; Carey, 1976; Benjamin, 1983; Macaulay *et al.*, 2005).

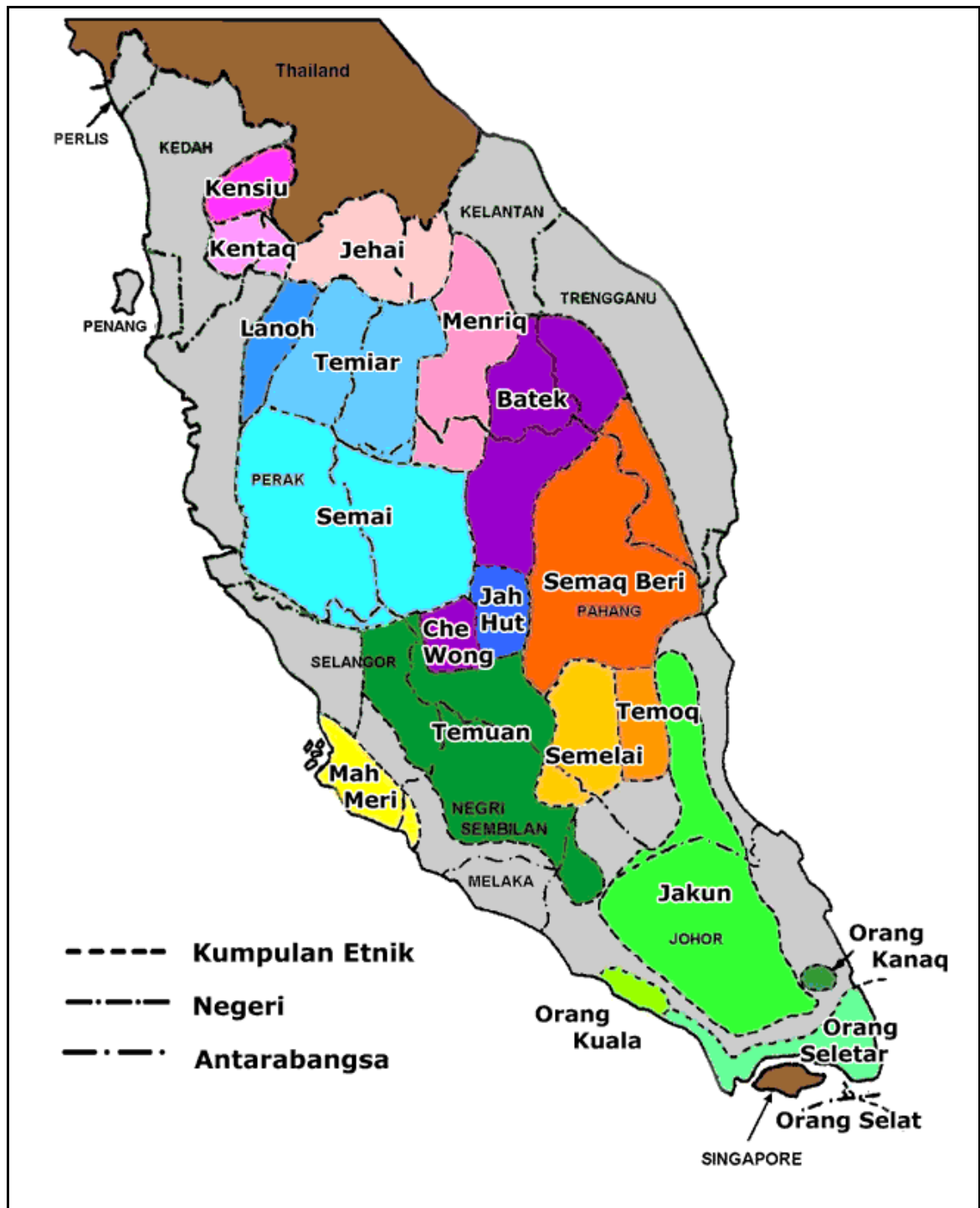
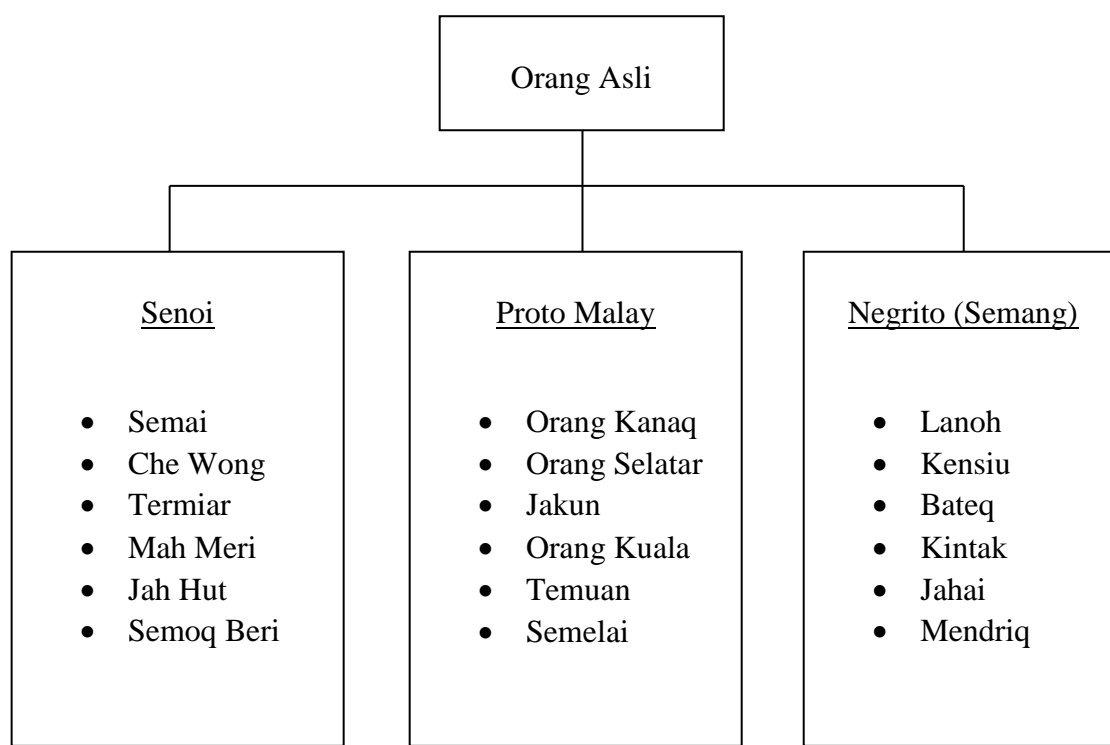


Figure 1.6: Map showing the Orang Asli distribution in Peninsular Malaysia

Table 1.1: Groups and sub-groups of Orang Asli in Peninsular Malaysia

(Source: <http://www.jakoa.gov.my/orang-asli/info-orang-asli/suku-kaumbangsa/>)



1.4.1(a) Semang

Semang also known as Negrito is the earliest Orang Asli group arrived in Peninsular Malaysia about 50,000 to 60,000 years ago (Macaulay *et al.*, 2005; Hill *et al.*, 2006). The Semang comprise six different sub-groups namely; Kensiu, Kintak, Jahai, Lanoh, Mendriq and Bateq. They have the least number of individuals among Orang Asli groups. Their settlements are mainly in the isolated area in the Northern and middle part of the Peninsular Malaysia (Baling Kedah, Hulu Perak, rural area of Kelantan, Terengganu and Pahang).

Semang are Austro-Asiatic speakers which are grouped with other Negrito communities in South Asia and Southeast Asia (SEA) such as Philippine Negritos, Mani in Thailand, Andaman islanders, and other phenotypically similar population in Papua New Guinea and Australia. These similarities have lead to the idea that all Negrito populations of SEA and Oceania originated from a common ancestral group which entered SEA during the earliest human dispersals into Asia (Endicott, 2013).

1.4.1(b) Senoi

Senoi is the largest group of Orang Asli in Peninsular Malaysia. They are divided into six sub-groups comprising the Semai, Che Wong, Temiar, Mah Meri, Jah Hut and Semoq Beri. The Senoi people inhabited the sloped of Titiwangsa namely in rural parts of Perak, Kelantan and Pahang. The term Senoi is derived from a Semai and Temiar word which mean people (Masron *et al.*, 2013).

It has been estimated that Senoi group reached Peninsular Malaysia during the second wave of migration about 8,000 years ago from the mountain areas of Cambodia and Vietnam (Baer, 1999). They have similar physical characteristics of Mongoloid and speak Khmer dialects but some believed that Senoi are descendants of Austroloid from Australia and Veddoid from South India (Fix, 1995).

1.4.1(c) Proto Malay

Proto Malay is the second largest group of Orang Asli and consist of six sub-groups; Temuan, Jakun/Orang Hulu, Seletar, Semelai, Kuala, and Orang Kanaq (Masron *et al.*, 2013). They reached Peninsular Malaysia later than the Senoi around 2,000 B.C. They were seafaring people and settled around central and coastal areas (Selangor, Negeri Sembilan, Melaka and Johor) of the Peninsular Malaysia (Bellwood, 2007).

Based on artifact evidence, linguistic and cultural, the Proto-Malays are people who migrated from the middle part of Asia (Yunnan) and came through Indo-China. Findings from archeology suggested that the proto-Austronesian speakers settled in Taiwan about 4,000 B.C. before migrated to Southeast Asia region through Philippines into Borneo, Sulawesi, Central Java and Eastern Indonesia around 2,500 years ago (Fix, 1995). From the morphological aspect, the Proto-Malay has similar morphology with Deutero-Malays (ancestor of Modern Malay) and share similarity in their culture and language (Kasimin, 1991).

In the present-day, Malays of the Malay Peninsular Malaysia are described as Deutero-Malays, who believed to be the descendants of the Proto-Malays admixed

with Thai, Indian, Siamese, Javanese, Sumatran, Chinese and Arab traders (Comas *et al.*, 1998). However, according to Fix (1995), the original Deutero-Malays migrated from southeast China (after migration of the Proto-Malays) over 1,500 years ago.

1.4.2 Population Genetics

Population genetics is the study of the variation in alleles and genotypes within the gene pool, and variation of changes from one generation to generation. Each member in the population receives its allele from their parents and passes them to their offspring. Based on developments in Mendel's laws of inheritance, the molecular understanding of genetics and modern evolutionary a mathematical technique are used predict the occurrence or combinations of specific alleles in populations (Kobilinsky *et al.*, 2005).

Allele frequency at a particular locus in a population can be calculated by sampling a number of individuals from that population. The allele frequencies may change depends on a few factors such as immigration or emigration, mutations, natural selection, non-random mating or the population being studied is relatively small in size causing genetic drift (Kobilinsky *et al.*, 2005). The Hardy-Weinberg equilibrium (HWE) describes and predicts a balanced in the frequencies of alleles and genotypes within a freely interbreeding population, assuming a large population size, no mutation, no genetic drift, no natural selection, no gene flow between population and random mating patterns. The stability of allele frequencies from generation to generation and Hardy-Weinberg Equilibrium in the population is important for forensic purposes (Budowle *et al.*, 2001).

Departure from HWE can be occurred when individuals within subpopulations have a tendency to mate with each other within the same group. Thus, form a small group of isolated individuals. National Research Council (NRC) committee on DNA technology has made a recommendation to apply a correction factor, known as theta (θ) value (0.03) in subpopulation calculations to acknowledge the subpopulations presentation in courtroom (Kobilinsky *et al.*, 2005).

1.4.3 Population Database

An allele frequency database is needed when evaluating the weight of evidence for an individual to be a contributor to a DNA sample. The allele frequency is required to access the genotype probabilities for unknown contributors of DNA to the sample. Typically databases are available from several populations all over the world. The common practice in DNA forensic field is to evaluate the weight of evidence using each available database for each unknown contributor. Sample size for each population in the database is must be more than 100 samples to make reliable (Chakraborty, 1992).

The main goal of having a population database is to find all common alleles and sample these alleles multiple times in order to estimate the frequency of alleles present in the population under consideration (Butler, 2010). The individuals that are sampled must come from homogenous group and data must be tested using the statistical model to ensure the allele frequencies are reasonable. With the assumption of independence in Hardy-Weinberg equilibrium and linkage equilibrium, it becomes possible to weigh the overall match probability.

1.5 Objectives of Study

General Objective:

The aim of this study is to understand the genetic history of the Orang Asli populations in Peninsular Malaysia using 15 autosomal STRs and 12 X-chromosomal STRs.

Specific Objectives:

- a) To analyze the genetic profile of 15 autosomal STRs and 12 X-chromosomal STRs for six Orang Asli sub-groups (Semai, Che Wong, Orang Kanaq, Lanoh, Bateq and Kensiu) in Peninsular Malaysia.
- b) To calculate the Hardy-Weinberg Equilibrium (HWE) and other forensic statistical parameters.
- c) To compare the genetic polymorphisms of the Orang Asli populations with other world populations.
- d) To develop autosomal STR and X-STR database for the Orang Asli populations in Peninsular Malaysia, particularly for forensic application.

CHAPTER 2

MATERIAL AND METHOD

2.1 Biological Samples

A total of 164 blood samples from six (6) Orang Asli sub-groups; Semai, Che Wong, Orang Kanaq, Lanoh, Bateq and Kensiu were collected. The sampling location and number of samples are shown in Table 2.1. The blood samples were collected after a brief interview and obtained informed consent (see Appendix A). To prevent the possibility of including Orang Asli individuals with admixed background in this study, very strict criteria was applied to ensure that there were no admixture with other sub-groups/ethnic for at least three generations (see Appendix B for Questionnaire). This study was approved by Universiti Sains Malaysia (USM) human ethics committee (see Appendix C).

Table 2.1: The sampling location and number of samples for six Orang Asli sub-groups

Orang Asli		Location	<i>N</i>
Group	Sub-group		
Senoï	Semai	Pos Tual, Kuala Lipis, Pahang	42
	Che Wong	Kampung Sungai Enggang, Lanchang, Pahang	28
Proto-Malay	Orang Kanaq	Kota Tinggi, Johor	11
Negrito (Semang)	Lanoh	Air Bah, Lenggong, Perak	26
	Bateq	Kampung Aring 5, Gua Musang, Kelantan	25
	Kensiu	Kampung Lubuk Legong, Baling, Kedah	32
TOTAL			164

Abbreviation: *N* = Number of samples.

The methodology for this study is summarized in the schematic diagram as shown below.

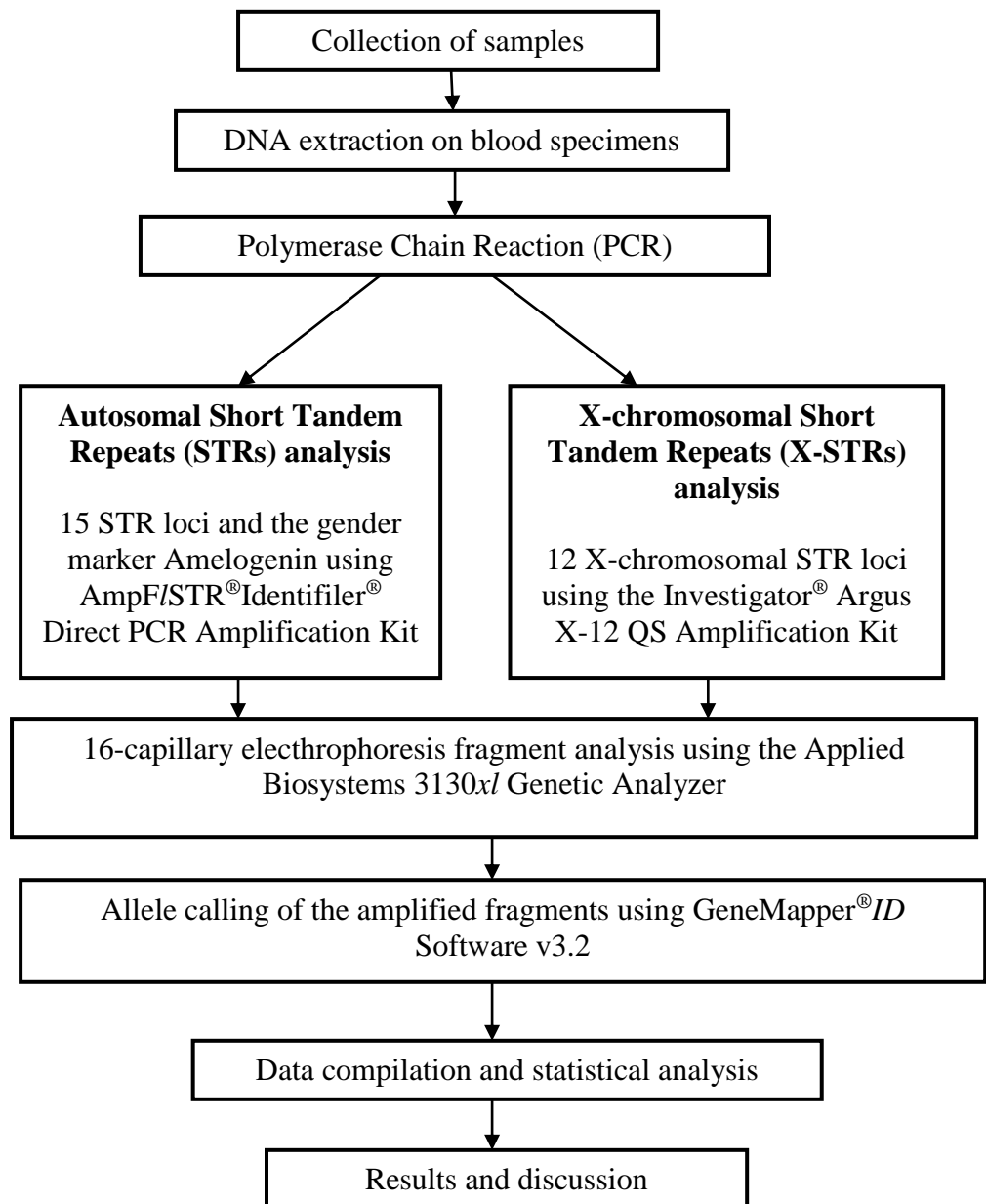


Figure 2.1: The schematic diagram of the methodology for the autosomal STR and X-STR analysis

2.2 Materials and Reagents

2.2.1 DNA Extraction Kit

For the DNA extraction, QIAamp[®] DNA Mini Kit was used. Each kit contains lysis buffer (Buffer AL), washing buffer (Buffer AW1 and AW2), elution buffer (Buffer AE) and Proteinase K for DNA extraction. Spin columns and collection tubes were also provided in this kit.

2.2.2 Human Polymerase Chain Reaction (PCR) Amplification Kit

2.2.2(a) AmpF/STR[®] Identifiler[®] Direct PCR Amplification Kit

This kit is a STR multiplex assay optimized for amplification of single-source DNA from blood and buccal samples (AmpF/STR Identifiler[®] Direct[®] PCR amplification kit user guide, 2015). The components of the kit are shown in Table 2.2. This kit uses combination of five-dye fluorescent system that enables direct amplification of the 15 autosomal STR loci, including D8S1179, D21S11, D7S820, CSF1PO, D3S1358, TH01, D13S317, D16S539, D2S1338, D19S433, vWA, TPOX, D18S51, D5S818, FGA and the sex-determining marker Amelogenin in a single PCR reaction. The amplified loci with their locations and corresponding dyes, the allelic ladder and the genotype of the control DNA 9947A are listed in Table 2.3.